4-2003

Recovery of Graded Response and Partial Credit Parameters in MULTILOG and PARSCALE

Christine E. DeMars
James Madison University, demarsce@jmu.edu

Running head: Nominal Response and Generalized Partial Credit

A Comparison of the Recovery of Parameters Using the Nominal Response and Generalized Partial Credit Models

Christine E. DeMars

James Madison University

(2004, April). Poster presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Abstract

In this simulation study, data were generated such that some items fit the generalized partial credit model (GPCM) while other items fit the nominal response model (NRM) but not the constraints of the GPCM. The purpose was to explore (a) how the errors in parameter estimation were affected by using the GPCM when the constraints of the GPCM were inappropriate, and (b) how the errors were affected by using the less-constrained NRM when the constraints of the GPCM were appropriate. With large sample sizes, there were considerable gains in precision from using the NRM when the GPCM was inappropriate, and only small losses in precision from using the NRM when the GPCM would have been appropriate. With small samples, there were greater benefits due to applying the constraints of the GPCM when appropriate, and smaller benefits due to using the NRM when the GPCM was inappropriate.

A Comparison of the Recovery of the Nominal Response and Generalized Partial Credit Models

Performance tasks and constructed response tests are often scored using polytomous rubrics, rubrics with multiple score points instead of simply right or wrong. Surveys of student and parent attitudes often employ polytomous response scales as well. Two possible item response models for polytomous data are the nominal response (Bock, 1972) and generalized partial credit models (Muraki, 1992). Both models describe the probability of responding or scoring in each of the item categories, given a respondent's trait score and the item's parameters. The nominal response model subsumes the generalized partial credit model as a constrained case. The purpose of these studies was to explore the consequences of using the more general model when the constrained model would fit the data, and of using the more constrained model in situations where the less-constrained model was somewhat more appropriate.

The generalized partial credit model (GPCM) is:

$$P_{ij}(\theta) = \frac{e^{a_i \sum_{k=1}^{j}(\theta - b_{ik})}}{\sum_{h=1}^{m_i} e^{a_i \sum_{k=1}^{h}(\theta - b_{ik})}} \qquad (1)$$

where

$P_{ij}(\theta)$ is the probability of scoring/selecting category j in item i, given trait score θ,

a_i is the item slope parameter,

b_{ij} is the category parameter (step difficulty) for category j (except for b_{i1}, which is arbitrarily defined as zero, b_{ij} is the transition where j - 1 and j are equally likely), and

m_i is the number of categories for item i.

Notice that this model differs from Masters' (1982) partial credit model in that each item can have a different *a*-parameter (slope); items with higher *a*-parameters are more discriminating. Within each

item, however, there is only one *a*-parameter. The option characteristic curves (OCCs) for a GPCM item, with $a = 1$ and $\mathbf{b} = [0, -0.75, -0.25, 0.25, 0.75]$ are diagrammed in Figure 1. For the GPCM, each *b* except the first is the intersection between adjacent categories. Notice that the OCCs for categories 1 and 5 are mirror images, and the middle categories have essentially the same shape except that they have different heights and different locations on the trait scale. Figure 1 also illustrates the role of the step parameters--the OCCs for categories 1 and 2 intersect at b_2 (-0.75), categories 2 and 3 intersect at b_3 (-0.25),etc.

The nominal response model (NRM) is:

$$P_{ij}(\theta) = \frac{e^{c_{ij} + a_{ij}\theta}}{\sum_{h=1}^{m_i} e^{c_{ih} + a_{ih}\theta}} \qquad (2),$$

where

$P_{ij}(\theta)$ is the probability of an examinee with a trait level of θ falling in category j of item i,

a_{ij} is the slope for category j of item i,

c_{ij} is the intercept for category j of item i, and

m_i is the number of categories of item i.

Bock (1972) did not give labels to the parameters a_{ij} and c_{ij}, simply calling them item parameters. The labels *slope* and *intercept* have been used by other researchers, including De Ayala and Sava-Bolesta (1999) and Wollack, Bolt, Cohen, and Young-Sun (2002). In contrast to the GPCM, the NRM has a different slope parameter for each category within the item. The response category that corresponds to highest levels of the trait will generally have the most positive *a*-parameter, while the category that corresponds to lowest levels of the trait will have the most negative *a*-parameter, but they are not necessarily mirror images; one may be almost flat while the other may be quite steep. Middle categories will have varying *a*-parameters, depending on whether

the probability of response more often decreases or more often increases as the trait level increases. For example, if the item described above and diagrammed in Figure 1 were written in terms of the NRM, the *a*-parameters would be [-2, -1, 0, 1, 2]. In this case, the highest and lowest categories have the same slope with reverse signs because the data fit the GPCM. The *a*-parameters for the middle categories are more difficult to interpret without graphing the function. An *a*-parameter of zero does not mean the category does not discriminate well the way an *a*-parameter of zero in the GPCM would mean the item had no discrimination. The *c*-parameters in Figure 1 would be [-0.5, 0.25, 0.5, 0.25, -0.5] in terms of the NRM model. Notice that the *c*-parameters are not the same as the intersection points of the categories the way they are in the GPCM, though these intersection points can be calculated based on the *a* and *c* parameters of adjacent categories. De Ayala and Sava-Bolesta (1999) explained that the *c*-parameters "reflect the interaction between a category's difficulty and how well it discriminates" (p. 4). The *c*-parameters are generally not easily interpretable in isolation, but they can be used in equation 2 to predict probability of response.

Figure 2 illustrates an NRM item that would not fit the GPCM; the *a*-parameters are [-1.5, -1.3, 0, .7, 2.5] and the *c*-parameters are [0.025, 0.175, 0.5, 0.325, -1.025]. The slopes for the item categories are not parallel or inverse, so these parameters could not be transformed into the GPCM parameters as those in Figure 1 could be. Also notice that the intersections between adjacent categories are the same as in Figure 1, but the *c*-parameters are completely different because the *c*-parameters depend not only on the intersection points but also the *a*-parameters of adjacent categories.

Another difference between the NRM and the GPCM is that the categories in the GPCM are in a pre-specified order; typically, the first category corresponds to the lowest level of the trait and successive categories correspond to higher levels of the trait. This does not necessarily mean that the *b*-parameters for a GPCM item are always in ascending order. For example, it is possible that

the intersection between categories 2 and 3 (b_3) occurs at a lower trait level than the intersection between categories 1 and 2 (b_2). However, even in this situation the probability of scoring 2 rather than 1, and the probability of scoring 3 rather than 2, would increase as the trait level increased. For the NRM model, the categories do not have to be in any particular order; the category labeled "1" in Figure 2 could have been designated "3" without affecting the characteristic curves.

In MULTILOG (Thissen, 1991), computer software for estimating item parameters for these and other polytomous models, the GPCM is specified as a special case of the NRM in terms of the matrices for the contrasts among the parameters. It is these contrasts that MULTILOG directly estimates, and the *a*-parameters and *c*-parameters can then be calculated from the contrasts. The default matrices are deviation matrices. For the GPCM, the *c*-parameter contrasts can be specified as triangular contrasts with the *a*-parameter contrasts as polynomial contrasts with all terms except the linear contrasts defined to be zero. The MULTILOG manual (Thissen, 1991, p. 2-21 and p. 3-20) as well as Childs and Chen (1999) give details on specifying these commands for the PCM; the only difference for the GPCM is that the *a*-parameters are not constrained to be equal across items.

The objective of this study was to compare the recovery of item parameters using the GPCM and NRM. The categories were in a meaningful order in all cases because the GPCM would not produce meaningful results otherwise. When the order of the item categories is known and a researcher or test developer believes a dataset will basically fit the GPCM, but some categories within an item may be somewhat more discriminating than others, it might seem sensible to use the NRM because if the data fit the GPCM, they should also fit the NRM, but the converse is not necessarily true. The test developer could fit both models and compare the difference in the -2 log-likelihood indices; a significant difference would indicate the NRM fit better, but this could mean that just a few of the items needed the less-constrained model. Many of the items might fit the more-constrained GPCM, which could only be discovered by changing the specifications for

individual items one at a time and comparing the difference in overall fit (the MULTILOG manual, Thissen, 1991, pp. 3-60 to 3-65, illustrates this approach). With long tests or multiple test forms, it is more likely that the same model will be used for all items, even when it is less than ideal. The drawbacks of using the NRM, if the GPCM does fit the data are: (1) the NRM has more parameters, so for any fixed sample size there is likely to be more error variance around the estimated parameters, (2) if the constraints of the GPCM are appropriate, relaxing those constraints could lead to chance overfitting of the model to small chance errors in the data, especially for small samples, and (3) conceptually, the item parameters for the NRM are less interpretable and harder to explain to others, which is particularly important to educational researchers who must release technical details to a broader audience. In this study, the data for half the items in each test were generated to fit the GPCM, and the data for the other half of the items were generated to fit the NRM; the response categories were ordered in both cases, but GPCM constraints on the a-parameters were not met for the NRM items. For the GPCM items, the errors of the item parameter estimates were expected to be smaller using the GPCM, but the question was whether the NRM errors would be meaningfully larger. For the items that fit the NRM but did not quite fit the GPCM, the NRM was expected to have smaller amounts of error, particularly with large datasets, though again the question was whether the error using the GPCM would be meaningfully larger. With small samples, the outcome was less predictable because of the possibility that the wrong, but more-constrained model (the GPCM) would lead to less error due to the difficulty of estimating the additional parameters in the NRM. For example, Lord (1980) suggested that with small sample sizes the one-parameter logistic (1PL) model might produce more accurate estimates than the three-parameter logistic (3PL) even when the 3PL was a better fit to the data. Similarly, Barnes and Wise (1991) found that constrained versions of the 3PL model, compared to the unconstrained 3PL model, led to smaller RMSEs in difficulty parameters and item characteristic curves.

Method

Four factors were studied: model used to generate the data (GPCM or NRM), model used to estimate the item parameters, number of items (20 or 10), and sample size of simulated respondents (2000, 500, or 250). All items had 5 response/score categories.

Data Simulation

For each of 100 replications, different sets of item parameters and trait parameters were used; for the 20 item condition, a total of 2000 items were simulated. All item parameters were initially selected in terms of the GPCM. The *a*-parameters were drawn from a log-normal distribution with mean -0.5, standard deviation 0.2. The first *b*-parameter for each item was arbitrarily set to zero and the second *b*-parameter (the first intersection point) for each item was drawn from a uniform distribution ranging from -2 to 1; the distance between each successive *b*-parameter and the previous one was then drawn from a uniform distribution ranging from 0.2 to 0.4. For the NRM parameters that did not meet the constraints of the GPCM, the parameters created for the GPCM were altered slightly. First, the GPCM item parameters were transformed into the equivalent NRM parameters. The GPCM *a*-parameter for each item was transformed to the equivalent NRM a-vector by multiplying the *a*-parameter by the linear contrast [-2, -1, 0, 1, 2]. Once the *a*-parameters were in terms of the NRM, they were altered so that the constraints of the GPCM would no longer fit. A constant of 0.5 was added to the first and last *a*-parameters, making the first category less discriminating and the last category more discriminating. For example, if the first *a*-parameter were -2 and the last *a*-parameter were 2, they would be altered to -1.5 and 2.5. Then 0.3 was subtracted from the second category and the fourth category, making the second category more discriminating and the fourth category less discriminating. The c-parameters were then calculated from the equation:

$$c_{k+1} = c_k - b_k(a_{k+1} - a_k) \tag{3}$$

where the *b*-parameters are the *b*-parameters, or intersections, from the GPCM. To deal with the indeterminacy in the location of the c's, c_1 was initially set to 0, then after solving Equation 3 for each of the other c-parameters, the *c*-parameters were rescaled by a constant such that they summed to zero. Figures 1 and 2, discussed above, illustrate an item before and after it was altered in this way. The reader can use these figures to judge the extent of the alteration. Note that the order of the categories and their intersection points were not changed; the *c*-parameters were altered simply because the intersections of the category functions are dependent on both the *a* and *c* vectors and the category intersections would have moved if the a-parameters were altered without adjusting the c parameters. A larger difference between the GPCM and NRM could have been created by changing the order of the categories, but realistically one would not try to fit the GPCM if one suspected the categories might not be ordinal.

For the test length of 20 items, for each replication 10 items were simulated under the GPCM and 10 under the NRM. For the test length of 10 items, 5 of the 10 GPCM and 5 of the 10 NRM items were randomly selected from each replication. Ten items would seem quite short for a multiple choice test, but would be realistic for a test with complex constructed-response items where students might spend 5-10 minutes on each item. It would also be realistic for one scale on a survey which measured multiple attitudes.

Trait scores for 2000 simulees were randomly selected for each replication from a normal (0,1) distribution. The conditions with sample sizes of 500 and 250 used subsets of these simulees. For the partial credit model (not Muraki's generalized version, but with slopes constrained to be equal across items), Choi, Cook, and Dodd (1997) found samples as small as 250 led to acceptable RMSEs for items with 4 categories, but larger samples of at least 1000 were needed for acceptable RMSEs for items with 7 categories. Five hundred has been suggested as a minimum for another polytomous model, the graded response model (Ankenmann & Stone, 1992; Reise & Yu, 1990).

The sample sizes for this study were chosen to span this range and included a large sample of 2000 that would be expected to have less error as well as a relatively small sample size of 250 which represented the lower end, or perhaps below the lower end, of the sample sizes that might be considered for polytomous model estimation.

Responses to each item were simulated by drawing randomly in proportion to the probability of the simulee choosing each category given the simulee's trait level and the item parameters. For example, if a simulee had probabilities of [0.10, 0.20, 0.25, 0.30, 0.15] of choosing/scoring [1, 2, 3, 4, 5] respectively, if a random draw from a uniform [0,1] distribution were between 0 and 0.1, the response would be coded 1; if between 0.1 and 0.3, the response would be coded 2; if between 0.3 and 0.55, the response would be coded 3, and so on.

Recovery of Parameters

MULTILOG (Thissen, 1991) was used to estimate the item parameters based on the simulated data. The item parameters were estimated twice, once using the unconstrained nominal model and once using the generalized partial credit model (triangle contrasts for the *c*-parameters, polytomous contrasts for the *a*-parameters, with all terms in the non-linear contrasts set to 0). Nineteen quadrature points were used for the normal trait distribution, evenly spaced from -4.5 to 4.5. Up to 300 cycles were specified to ensure convergence in all conditions. Otherwise, program defaults were used.

Two measures were used to assess the accuracy of the recovered item parameters: the root-mean-square-error (RMSE) of the parameters, and the RMSE between the OCC's. The RMSEs are the empirical standard errors (the standard deviation around the true parameter, not around the average estimate, such that the RMSE includes any bias as well as variance in the parameter estimation). The smaller the RMSE, the more precise the estimate is, on average. The parameters

were left in terms of the NRM because, while GPCM parameters can be transformed to NRM parameters, NRM parameters generally can not be transformed to GPCM parameters.

The RMSE of the *a*-parameters or *c*-parameters was calculated within each condition as:

$$\mathrm{RMSE}_{\Lambda} = \sqrt{\frac{\sum_{h=1}^{r}\sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{\Lambda}_{hij} - \Lambda_{hij})^2}{rnm}}, \tag{4}$$

where

Λ_{hij} is the parameter (*a*- or *c*-parameter) for category j in item i in replication h

$\hat{\Lambda}_{hij}$ is the estimate of Λ_{hij},

m is the number of response/score categories,

n is the number of items, and

r is the number of replications.

For each OCC, the RMSE between the true and recovered curve was approximated by evaluating the functions (true and recovered) at each quadrature point, squaring the difference between the probabilities, weighting by the density at the point, and summing across the quadrature points. This can be symbolized:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{h=1}^{r}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{q=1}^{Q} w_q[\hat{P}_{hij}(\theta_q) - P_{hij}(\theta_q)]^2}{rnm}}, \tag{5}$$

where

w_q is the density at quadrature node q in a normal distribution,

$\hat{P}_{hij}(\theta_q)$ is the probability of a respondent at node q responding in category j of item i (in replication h), based on the estimated parameters, and

$P_{hij}(\theta_q)$ is the probability of a respondent at node q responding in category j of item i (in replication h), based on the true parameters.

Similar, but not identical, indices were used by Hulin, Lissak, & Drasgow (1982) and Drasgow (1989) in studies of the recovery of parameters for dichotomous items. In this study, greater weight was given at the center of the latent trait distribution, where the greatest number of examinees or respondents would be.

θ's were estimated by maximum likelihood (ML). The bias (average difference between the estimated and true value) and RMSE were calculated for the θ estimates as well. Based on their true θ's, simulees were grouped into intervals of width = .25 and bias and RMSE were estimated within each interval. Test length would be expected to have a direct effect on the RMSE of the θ's, as each additional item should decrease the standard error. Sample size could have an indirect effect; sample size would be expected to influence the accuracy of the item parameter estimation, which in turn could impact the RMSE of the θ's.

Results

Overall Fit of the Models

The difference between the -2 log-likelihood for the GPCM and the -2 log-likelihood for the NRM was used to test whether the NRM fit significantly better than the GPCM. This test is meaningful here because, if the difference were not significant, in the interest of parsimony one would generally use GPCM. If the difference were significant, one could rerun both models, relaxing the constraints one item at a time to find which items did not meet the GPCM, or one could use the NRM for all items. The difference in -2 log-likelihoods, under the null model of no difference in fit, would be distributed as chi-square with 60 degrees of freedom for the 20-item tests and 30 degrees of freedom for the 10-item tests. This difference should have been significant in all

conditions, because half of the items in each test did not meet the constraints of the GPCM. The difference was statistically significant at the 0.05 level in 100% of the replications of all conditions except the 10-item test with sample size of 250, where the difference was significant in 98 of the 100 replications.

RMSE of Parameters and OCCs

RMSEs are shown, by condition, in Figures 3-8. In Figures 3 and 4, it appears test length had little impact on the RMSEs of the a-parameters when the GPCM was used to estimate the items parameters. When the NRM was used for estimation, RMSEs were somewhat smaller when the test was longer. When the data were generated to follow the GPCM (Figure 3), the RMSEs were smaller when the GPCM, rather than the less-constrained NRM, was used to recover the item parameters. This difference noticeably decreased as the sample size increased. When the data were generated to follow the NRM but not the more-constrained GPCM (Figure 4), the RMSEs were smaller when the NRM, rather than the GPCM, was used to recover the item parameters. This difference was greatest when the sample size was largest; when the sample size was 250, the advantage of the NRM was small for the 20-item test and actually reversed for the 10-item test.

In Figures 5 and 6, the results followed the same pattern for the c-parameters as shown in Figures 3 and 4 for the a-parameters. The RMSEs were slightly larger for the c-parameters, but the effects of the factors were the same.

The RMSEs for the OCCs are illustrated in Figures 7 and 8. These RMSEs show how well the probability functions, as opposed to the individual parameters, were reproduced. Test length had no meaningful impact on the RMSEs. When the data were generated to follow the GPCM (Figure 7), the RMSE was again smaller when the GPCM was used for parameter estimation with smaller samples, but this difference was smaller than it was for the individual parameters, and it virtually disappeared for the sample size of 2000. When the data were generated to follow the less-

constrained NRM (Figure 8), the RMSE was smaller when the NRM was used for parameter recovery, as was the case for the a- and c-parameters. The advantage of the NRM over the GPCM decreased as the sample size decreased, but the RMSE remained larger for the GPCM estimation even with the sample size of 250.

Bias and RMSE of θs

Bias in θ estimation is shown in Figures 9 and 10. The θ estimates were based on the total test (all 10 or 20 items); half the items fit the GPCM and half did not, so when the GPCM was used it was inappropriate for half the items and when the NRM was used it was unnecessarily complicated for half the items. For the NRM, the bias showed the typical pattern of ML estimates; the more extreme θ's were biased outwards. For the GPCM, the bias showed an unexpected pattern; estimates at *both* extremes were higher than their true values. Also, the bias at the center was slightly negative. The absolute value of the bias was 0.1 or lower for θ's from about -1.5 to 2 for the NRM and from about -1.5 to 1 for the GPCM. Outside of this range, bias tended to be greater for the GPCM.

RMSE for θ (Figures 11 and 12) was similar for the NRM and GPCM for θ's from about -1.5 and higher. For lower θ's, though, RMSE was noticeably lower for the GPCM, despite the fact that bias was higher for the GPCM in this range. Test length made a difference; under both models, there were two noticeable bands of RMSEs, with the 20-item tests resulting in smaller RMSE. Test length had a much bigger impact than the number of examinees used to calibrate the items.

Limitations and Conclusions

Like all simulation studies, this study was limited by the specific conditions chosen. Only a limited number of sample sizes and test lengths were studied; test length might have had an impact if very short tests were included. Only normally distributed latent traits were studied; departures from normality could potentially influence RMSE; results with other IRT models have been mixed

though any differences due to non-normality have generally been small (Kirisci, Hsu, & Yu, 2001; Reise & Yu, 1990; Stone, 1992; Swaminathan & Gifford, 1983). Most importantly, the degree to which the data departed from the GPCM likely had a strong influence on the results. The purpose was to investigate obvious departures from the GPCM in terms of category slopes (as evidenced by comparing Figures 1 and 2) while still keeping the categories in a fixed order so that the GPCM was not completely unreasonable. Smaller departures from the model would have had less of an effect.

The inclusion of different randomly selected items in each replication also presented limitations. This procedure was selected for greater generalizability; chance combinations of item parameters which were particularly hard to estimate would not recur over and over again in every replication and thus would not unduly influence the results. However, this also meant that these cases could not be identified and classified because RMSE could not be calculated for specific parameters.

A final limitation was that each dataset had some items that followed the GPCM and some that did not. The models were mixed within datasets because it is in this situation where one is most likely to want to use the less optimal model for some items, for the simplicity of using the same model for all the items. If the test of fit showed that the NRM fit better overall, it would be simplest to use the NRM for all items. However, this mixing of models made it impossible to test the Type I error rate as each test always included some items which did not fit the GPCM.

When large samples (n = 2000) were used, if the data did not follow the GPCM the RMSEs for the item parameters and characteristic curves were considerably smaller when the parameters were estimated using the NRM. If the data did follow the GPCM, RMSEs for the individual parameters were somewhat smaller when the GPCM was used to estimate the parameters, but the RMSEs for the characteristic curves were about the same regardless of which estimation model was utilized. This makes sense because the NRM is not an inappropriate model when the GPCM fits, it

simply has more free parameters than necessary and is estimated less optimally. Therefore, one could reasonably run the NRM knowing that little precision would be lost for items where the GPCM would have been appropriate, and precision would be gained for items where the GPCM was not appropriate. Particularly where the comparison of the -2 loglikelihoods from the GPCM and NRM indicated that overall the fit was better to the NRM, not much precision would be lost by then using the NRM for all items.

With smaller sample sizes, when the data followed the GPCM the advantage (in relative precision) of using the GPCM over the NRM to estimate the item parameters increased compared to the larger samples. This was more true for the errors of the parameter estimates than for the errors of the OCCs. When the data departed from the GPCM, the relative advantages of using the correct NRM for estimation were less for the smaller samples than they were for the larger samples. With smaller samples, there is more to gain by using the GPCM when it is appropriate, and relatively less to lose if it is not. However, if viewed in an absolute rather than relative sense, the RMSEs for the NRM data with sample size of 250 were large regardless of which model was used in parameter recovery. For this sample size, it might be best to test whether the NRM offered significantly better fit than the GPCM, and if so to avoiding releasing or using item parameter estimates until more data can be collected.

When estimating θ using these tests where half the items fit the GPCM but half did not, for θ's below -1 the RMSEs were smaller using the GPCM. This was due not to an decrease in bias using the GPCM but to less variance. Longer tests, as would be expected, led to lower levels of RMSE. Sample size had little impact on precision of θ estimates; samples as small as 250 produced θ estimates that were about as accurate as the calibrations from larger samples. However, accurate item parameterization is important for equating, DIF studies, or test development, so in most contexts larger samples are still needed.

References

Ankenmann, R. D., & Stone, C. A. (1992, April). *A Monte Carlo study of marginal maximum likelihood parameter estimates for the graded model.* Paper presented at the annual meeting of the National Council for Measurement in Education, San Francisco, CA. (ERIC Document Reproduction Service No. ED347189)

Barnes, L. L. B., & Wise, S. L. (1991). The utility of a modified one-parameter IRT model with small samples. *Applied Measurement in Education, 4*, 143-157.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Childs, R. A., & Chen, W.-H. (1999). Software note: Obtaining comparable item parameter estimates in MULTILOG and PARSCALE for two polytomous IRT models. *Applied Psychological Measurement, 21*, 89-90.

Choi, S. W., Cook, K. F., & Dodd, B. G. (1997). Parameter recovery for the partial credit model using MULTILOG. *Journal of Outcome Measurement, 1*, 114-142.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13*, 77-90.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*, 249-260.

Kirisci, L., Hsu, T. C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*, 146-162.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133-144.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*, 1-16.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New Horizons in Testing: Latent trait test theory and computerized adaptive testing* (pp. 13-30). New York: Academic Press.

Thissen, D. (1991). MULTILOG: Multiple, categorical item analysis and test scoring using item response theory [Computer software]. Chicago: Scientific Software.

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y.-S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 26*, 339-352.

Figure Captions

Figure 1. A generalized partial credit model item.

Figure 2. A nominal response model item.

Figure 3. RMSE of *a*-parameters, data simulated under the GPCM.

Figure 4. RMSE of *a*-parameters, data simulated under the NRM.

Figure 5. RMSE of *c*-parameters, data simulated under the GPCM.

Figure 6. RMSE of *c*-parameters, data simulated under the NRM.

Figure 7. RMSE of option characteristic curves, data simulated under the GPCM.

Figure 8. RMSE of option characteristic curves, data simulated under the NRM.

Figure 9. Bias in θ estimated using GPCM item parameters.

Figure 10. Bias in θ estimated using NR item parameters.

Figure 11. RMSE in θ estimated using GPCM item parameters.

Figure 12. RMSE in θ estimated using NR item parameters.

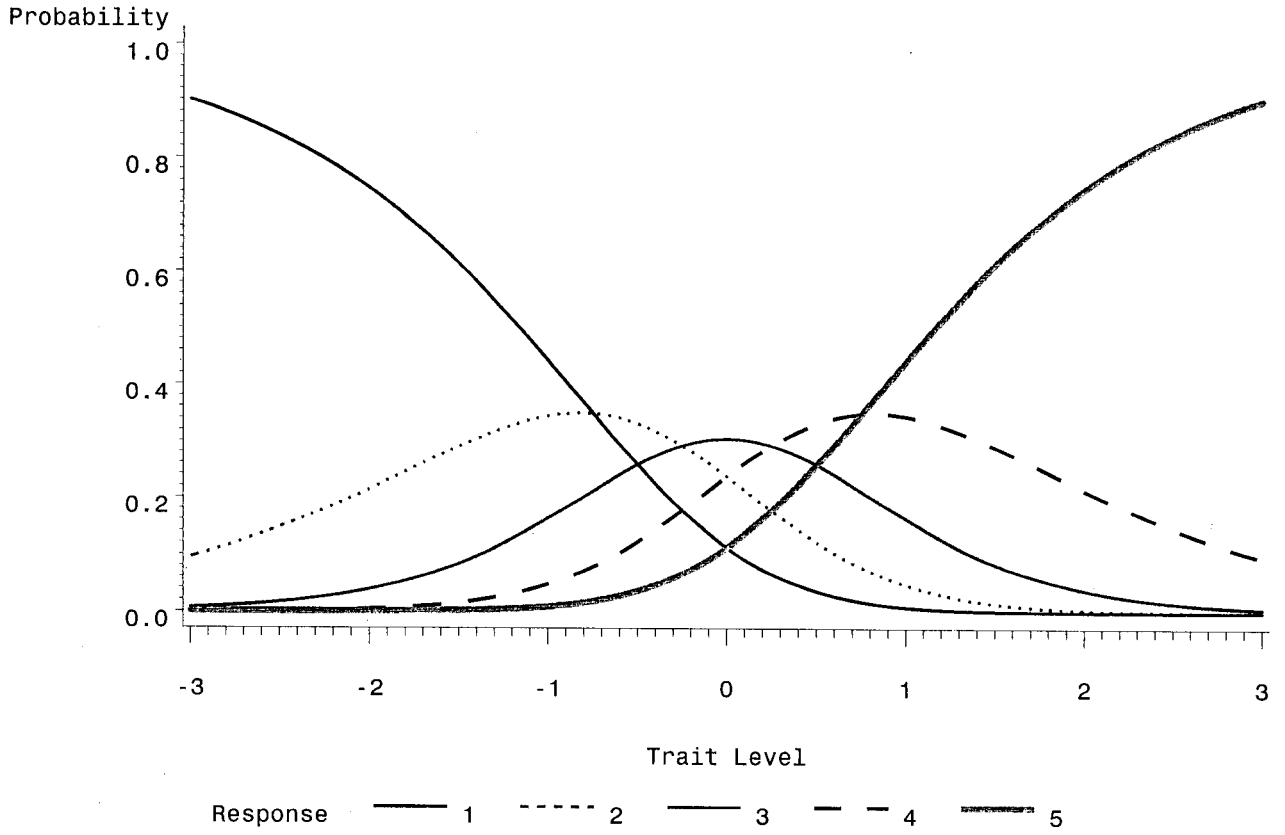


Figure 1. A generalized partial credit model item.

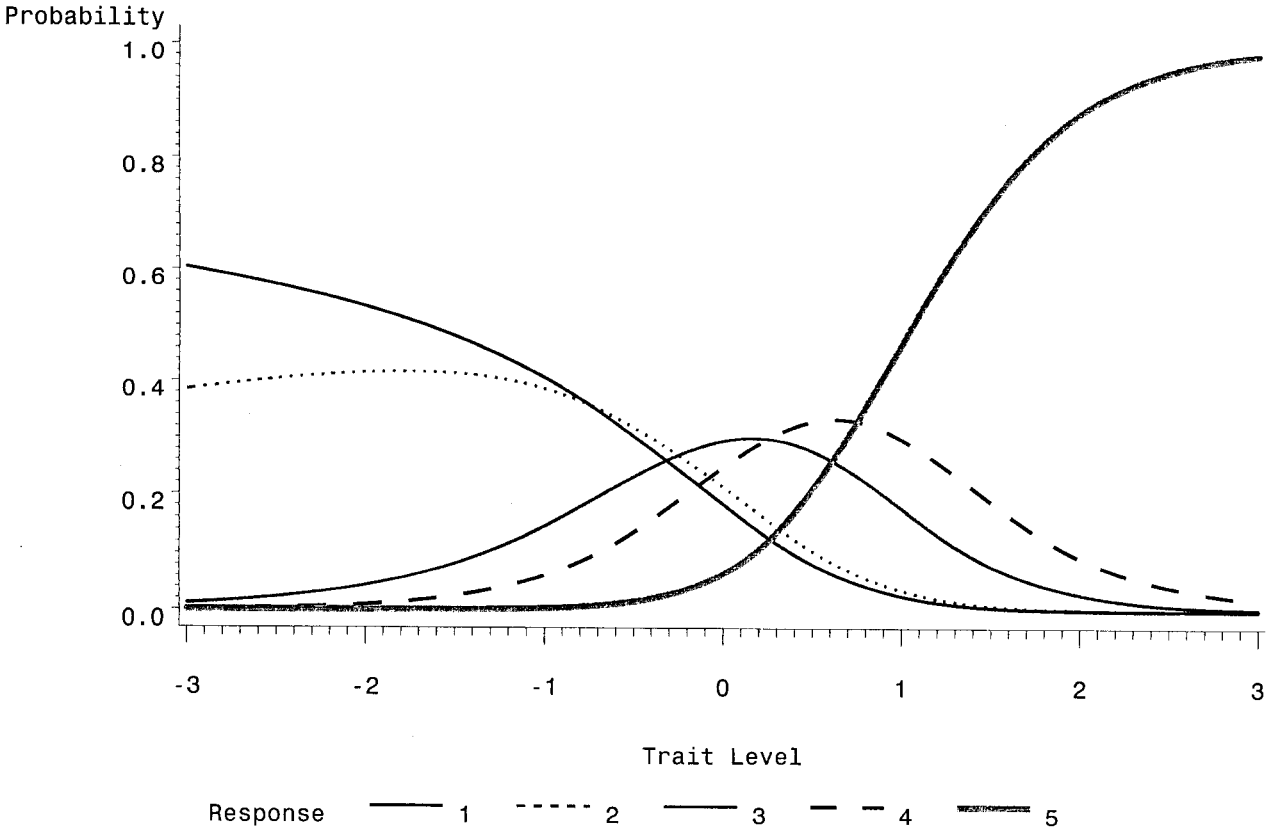
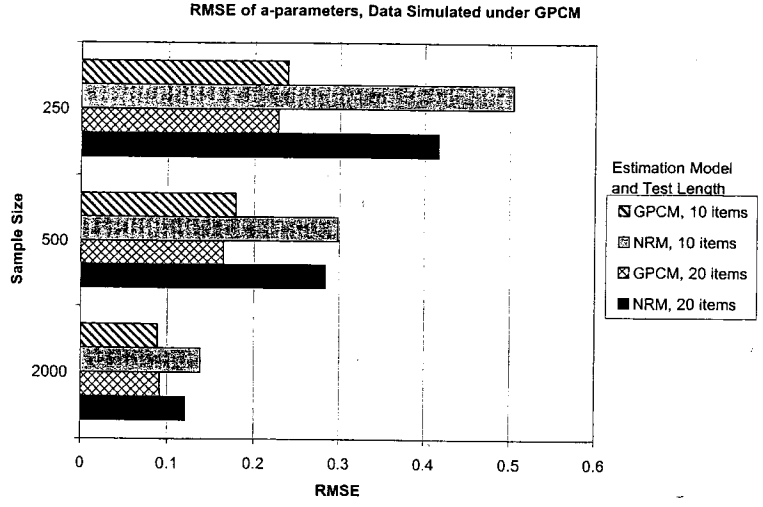
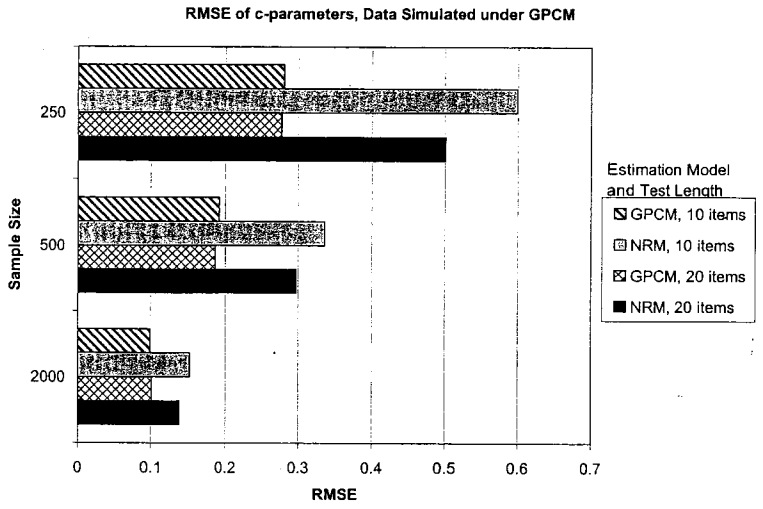
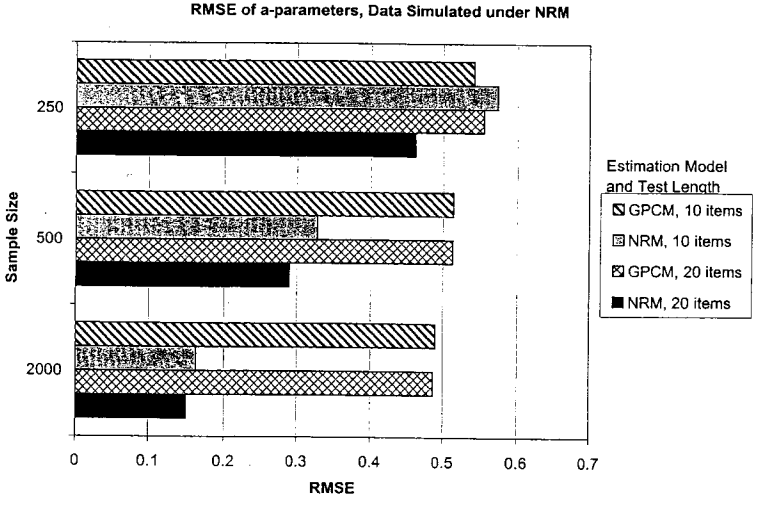
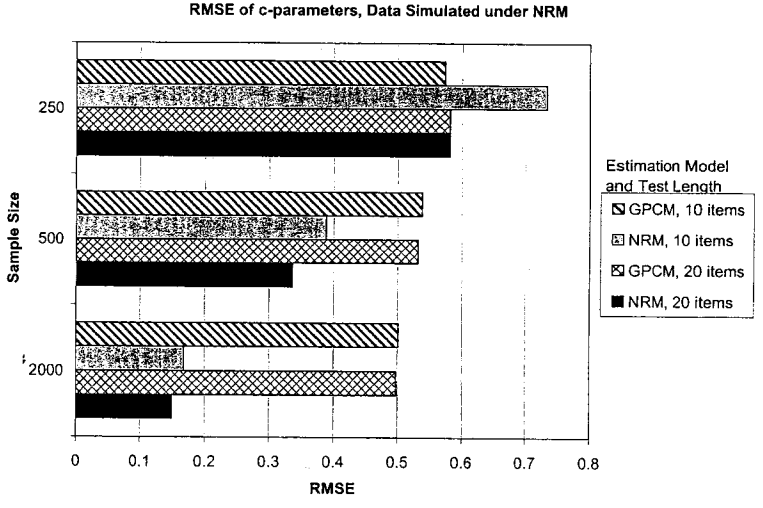


Figure 2. A nominal response model item.

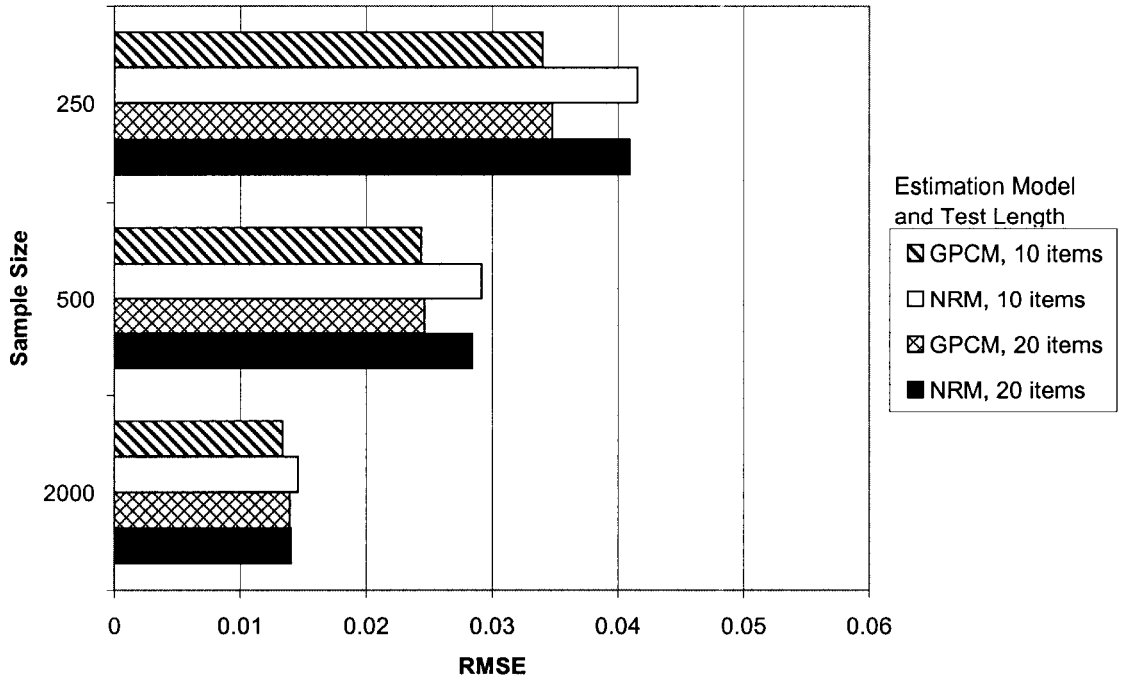

RMSE of a-parameters, Data Simulated under GPCM
Sample Size
250
500
2000
0
0.1
0.2
0.3
0.4
0.5
0.6
RMSE
Estimation Model and Test Length
GPCM, 10 items
NRM, 10 items
GPCM, 20 items
NRM, 20 items


RMSE of c-parameters, Data Simulated under GPCM
Sample Size
250
500
2000
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
RMSE
Estimation Model and Test Length
GPCM, 10 items
NRM, 10 items
GPCM, 20 items
NRM, 20 items

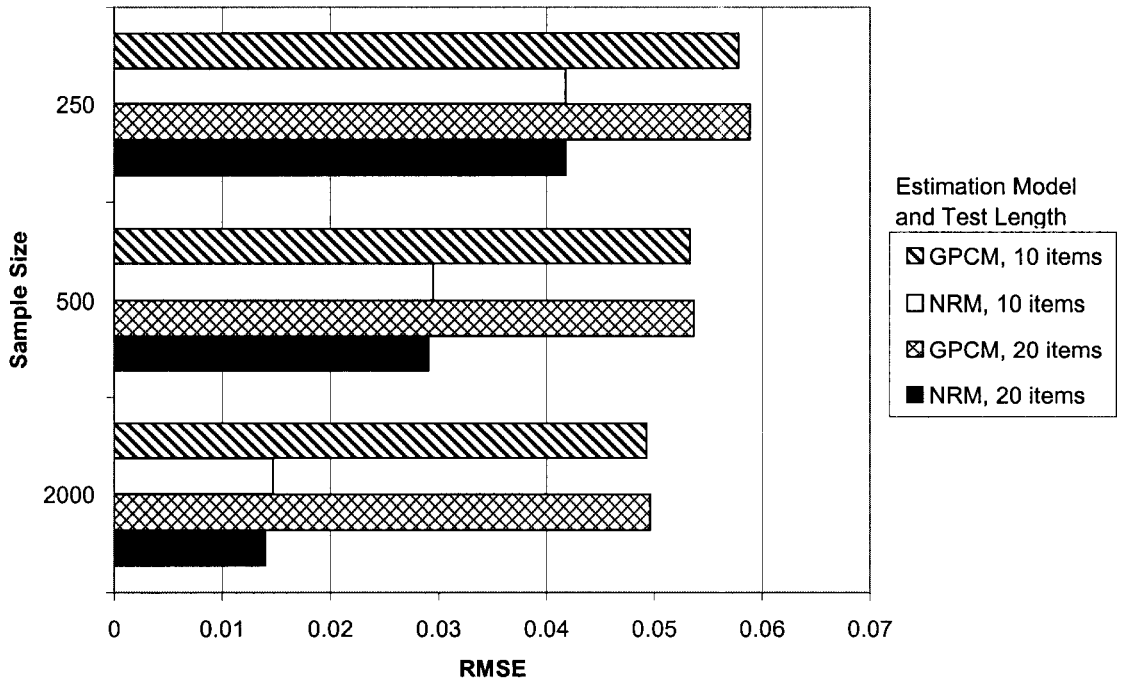

RMSE of a-parameters, Data Simulated under NRM
Sample Size
250
500
2000
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
RMSE
Estimation Model and Test Length
GPCM, 10 items
NRM, 10 items
GPCM, 20 items
NRM, 20 items


RMSE of c-parameters, Data Simulated under NRM
Sample Size
250
500
2000
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
RMSE
Estimation Model and Test Length
GPCM, 10 items
NRM, 10 items
GPCM, 20 items
NRM, 20 items

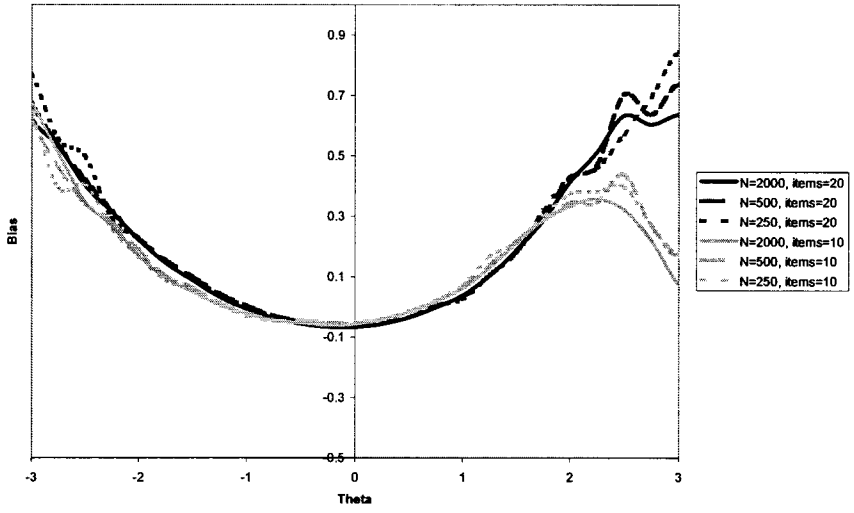
RMSE of Option Characteristic Curves, Data Simulated under GPCM



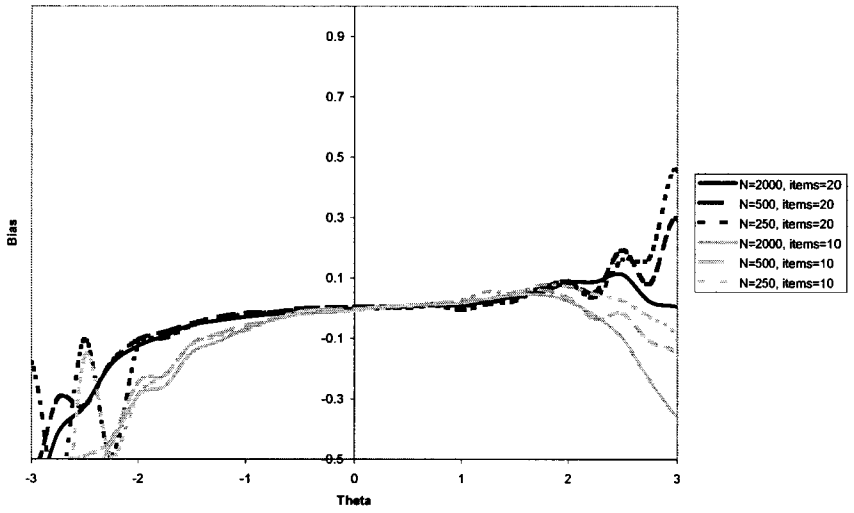
RMSE of Option Characteristic Curves, Data Simulated under NRM



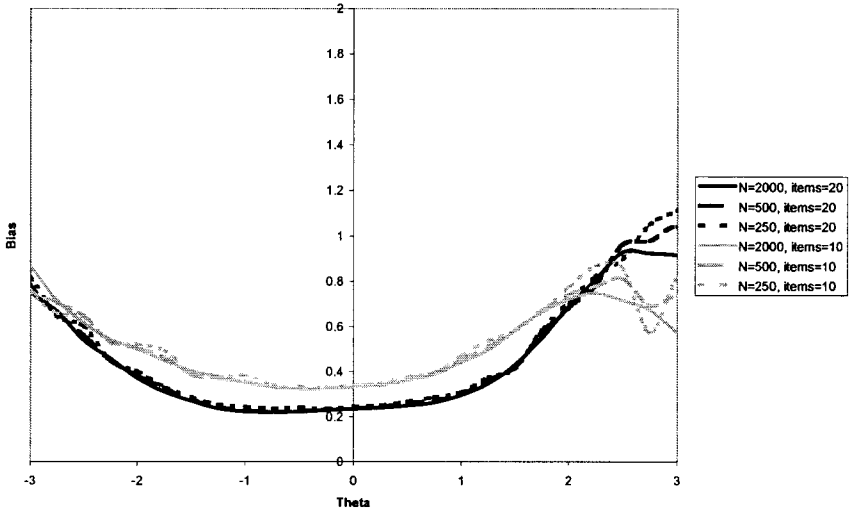
Bias in θ using Generalized Partial Credit Model



Bias in θ using Nominal Response Model



RMSE in θ using Generalized Partial Credit Model



RMSE in θ using Nominal Response Model